Statistical methods for analysing complex genetic traits

El Galta, Rachid

CHAPTER 6

Generalizing Terwilliger's likelihood approach: a new score statistic to test for genetic association

R. el Galta, S. Uitte de Willige, M.C.H. de Visser, L. Hsu, J.J. Houwing-Duistermaat

Abstract

In this report, we propose a one degree of freedom test for association between candidate genes and binary traits. We consider the situation of observed haplotypes, i.e. haplotype tagging single nucleotide polymorphisms are typed from which the haplotypes can be derived with almost 100% certainty. The method is a generalization of the approach of Terwilliger (1995) and is especially powerful for the situation of one associated haplotype. As alternative to the likelihood ratio statistic, we derive a score statistic, which is locally most powerful. By means of a simulation study, we compare the performance of the score statistic to Pearson's chi-square statistic and the likelihood ratio statistic proposed by Terwilliger (1995). We illustrate the method on three candidate genes studied in the Leiden Thrombophilia Study (Uitte de Willige et al., 2005). We conclude that the statistic follows a chi square distribution under the null hypothesis and that the score statistic has often more power than Terwilliger's likelihood ratio statistic, especially for variants with frequencies between 0.1 and 0.4 and which have a small impact on the studied disorder.

Submitted for publication

6.1 Introduction

For genetic association studies, single nucleotide polymorphisms (SNPs) are popular genetic markers, because they are stable, easier to type than the micro satellite markers and distributed with a high density over the genome (Collins et al., 1997). The pattern of variants appears to be structured with sets of physically close SNPs inherited together in blocks (Daly et al., 2001; Gabriel et al., 2002). Within a block, SNPs are highly correlated (linkage disequilibrium) and each block contains only a few common haplotypes (Gabriel et al., 2002). These haplotypes can be described by a small number of SNPs. Several methods are described to identify the minimal informative subset of SNPs, so called haplotype tagging SNPs (htSNPs) (see Sebastiani et al. (2003) and Stram (2004) for references).

In the case of zero recombination within a block, the haplotypes can be uniquely identified and *n* haplotypes can be described by *n-1* SNPs (Bafna et al., 2003; Clark, 2004; Clayton et al., 2004). Stram (2004) introduced a measure for reliability of haplotype assignment r_h^2 (Epstein and Satten, 2003; Satten and Epstein, 2004). For r_h^2 close to one, the haplotypes are known and association between the observed haplotypes and a disease can be studied by comparing the haplotype frequencies of the gene in cases and controls. Examples in the literature of genes with known haplotypes are APOE (Fullerton et al., 2000), CRP (Carlson et al., 2005) and the fibrinogen gamma (FGG), fibrinogen alpha (FGA), and fibrinogen beta (FGB) genes (Uitte de Willige et al., 2005). Note that Carlson et al. (2005)) used the haplo.stats package (Schaid et al., 2002) which does take into account the uncertainty in phase by using the EM algorithm and then they noted that the same results could be obtained by using the htSNPs. The reason for this is that the htSNPs uniquely correspond to the haplotypes.

If a causal mutation occurred in one haplotype in the past, it would be natural to consider haplotypes rather than individual genotypes (Clark, 2004) and to assume that only one haplotype carries the causal variant (Terwilliger, 1995). The haplotype frequencies in cases can therefore be modelled by the frequencies in controls plus one additional parameter, which accounts for the excessiveness of the causal haplotype. Under this assumption, a statistic which tests the null hypothesis of this additional parameter equal to zero will have more power than the classical chi-square test, since the latter tests for any differences in frequencies between cases and controls.

Terwilliger (1995) proposed this model in the context of multi-allelic mark-

ers and used the likelihood ratio test, i.e. comparing the likelihood under the alternative to the likelihood under no association, for testing. Since it is unknown which marker allele is associated with the disease, the likelihood corresponding to this model is a weighted sum over all alleles i of conditional likelihoods given that allele i is over-represented in the set of cases. As for weights Terwilliger (1995) proposed to use the allele frequencies in the population from which the cases and controls were sampled. The excess frequency of the associated allele in cases is modelled by the parameter λ which is the fraction attributable at risk (Clayton, 2000). The corresponding log likelihood function has a number of unusual features. For example, the allele frequencies that are used as weights are unknown parameters in the conditional likelihood functions. The score function, the first derivative of the log likelihood function with respect to λ evaluated at $\lambda = 0$ is a constant zero for any observed data. Therefore, the distribution of the likelihood ratio (TLR) statistic under the null hypothesis is not straightforward and the $50 : 50$ mixture of chi square distributions of null and one degrees of freedom, which was suggested by Terwilliger (1995), appears to yield conservative p-values (Sham et al., 1996).

Under Hardy-Weinberg equilibrium and complete linkage disequilibrium, the observed haplotype frequencies in the controls agree with the population frequencies many generations ago. Hence the frequency of a haplotype in the population corresponds to the prior probability that the mutation occurred on that haplotype. If one is focussed in detecting common haplotypes with a small impact on the trait, an alternative for the haplotype frequencies is a flat prior. The advantages of using a flat prior are that the probabilities do not have to be estimated. Furthermore the first derivative of the log likelihood with respect to λ evaluated at $\lambda = 0$ is not equal to zero and the score statistic as alternative for the likelihood ratio statistic can be used. Advantages of score statistics compared to likelihood ratio statistics are that they are also locally most powerful, and because they do not need to evaluate the log likelihood under the alternative, they are often easier to compute and robust to small model deviations under the alternative (Cox and Hinkley, 1974).

In this paper we consider the score statistic as alternative to the classical chi-square and the original TLR statistic of Terwilliger (1995) in the context of haplotypes. We also include the likelihood ratio statistic corresponding to the log likelihood using equal weights. We carried out a simulation to study the performance of the four statistics under the null hypothesis and to compare the power of the four statistics under various alternatives. Finally we illustrate the proposed statistics by an association analysis of three candidate genes in

the Leiden Thrombophilia Study (LETS) (Uitte de Willige et al., 2005).

6.2 Methods

Let m be the number of haplotypes describing most of the genetic variation in a gene. Assume that the haplotype frequencies are in Hardy-Weinberg equilibrium proportions. Let $p = (p_1, \cdots, p_m)$ be the vector of haplotype frequencies in controls. Assume that only one haplotype denoted with index i is over-represented in the cases, then the haplotype frequencies in the cases can be modelled as $q_i = p_i + \lambda(1 - p_i)$ and $q_j = p_j - \lambda p_j$ for $j \in (1, \cdots, i-1, i+1, \cdots, m)$. Let $x = (x_1, \cdots, x_m)$ and $y = (y_1, \cdots, y_m)$ be vectors of haplotype counts in the cases and the controls, respectively, and let n_1 and n_2 be the total number of case chromosomes and of control chromosomes, respectively, and $n = n_1 + n_2$. Then the conditional likelihood L_i given that haplotype i carries the mutation is given by

$$L_i(\lambda, p|x, y) = (p_i + \lambda(1 - p_i))^{x_i}(1 - \lambda)^{n_1 - x_i} \prod_{j \neq i}^{m} p_j^{x_j} \prod_{j=1}^{m} p_j^{y_j} \tag{6.1}$$

and the likelihood proposed by Terwilliger is equal to

$$L(\lambda, p|x, y) = \sum_{j=1}^{m} p_j L_j, \tag{6.2}$$

with L_j given in formula (6.1).

It is easy to see that likelihood function (6.2) can be generalized to the following likelihood function:

$$L(\lambda, p|x, y, w) = \sum_{j=1}^{m} w_j L_j,$$

with L_j given in formula (6.1) and $w = (w_1, \cdots, w_m)$ a vector of known positive weights restricted by $\sum_{j=1}^{m} w_j = 1$. The first derivative of the log likelihood $l(\lambda, p|x, y, w) = log(L(\lambda, p|x, y, w))$ to λ evaluated in $\lambda = 0$ is equal to

$$\begin{aligned} U_w &= \frac{\partial}{\partial \lambda} l(\lambda, p|x, y, w)_{|\lambda=0} \\ &= \sum_{j=1}^{m} \frac{w_j(x_j - n_1 p_j)}{p_j}. \end{aligned}$$

For known allele frequencies p_j, the distribution of the U_w under H_0 can be approximated by the normal distribution with zero mean and variance $\text{VAR}[U_w] = n_1(\sum_{j=1}^{m} w_j^2 p_j^{-1} - 1)$. Note that $U_w = 0$ when for all j $w_j = p_j$.

Often the haplotype frequencies are unknown and have to be estimated from the data. Under the null hypothesis we can estimate the frequencies from the combined sample of cases and controls by $\hat{p}_j = \frac{x_j + y_j}{n}$ and an estimate of the score statistic U_w is given by

$$\hat{U}_w = \sum_{j=1}^{m} \frac{w_j(x_j - n_1\hat{p}_j)}{\hat{p}_j}.$$

under H_0, $\hat{U}_w$ has approximately mean equal to zero and variance

$$\text{VAR}(\hat{U}_w) \approx n^{-1} n_1 n_2 (\sum_{j=1}^{m} w_j^2 p_j^{-1} - 1). \tag{6.3}$$

Note that the variance $\text{VAR}(\hat{U}_w)$ is increased by n_2/n fold compared to the variance $\text{VAR}(U_w)$ because the allele frequencies are estimated from the data. Now the score statistic $\hat{S}_w$ is defined by

$$\hat{S}_w = \frac{\hat{U}_w^2}{\hat{\text{VAR}}(\hat{U}_w)},$$

where $\hat{\text{VAR}}(\hat{U}_w)$ is obtained by replacing p_j by its estimate $\hat{p}_j$ in formula (6.3). When all haplotypes are common, a natural choice of weights is $w_j = \frac{1}{m}$.

Under the alternative hypothesis of the presence of one positively associated haplotype i, the expectation of $\hat{U}_{\frac{1}{m}}$ is

$$E_{H_A}[\hat{U}_{\frac{1}{m}}] \approx \frac{n_1 n_2 \lambda}{n - n_1\lambda}\left(\frac{n}{n_1 q_i + n_2 p_i} - m\right), \tag{6.4}$$

with $\frac{n_1 p_i + n_2 q_i}{n}$ the frequency of the associated haplotype in the combined sample and $q_i = p_i + \lambda(1 - p_i)$. When $\frac{n_1 p_i + n_2 q_i}{n}$ is larger than $\frac{1}{m}$ this expectation becomes negative. Therefore we propose a chi-square distribution with one degree of freedom to approximated the distribution of this statistic under the null hypothesis.

6.3 Results

Simulation study

By means of a simulation study, we first evaluated the type I error rate of the score statistic $\hat{S}_{\frac{1}{m}}$, Pearson's chi square χ^2, the likelihood ratio with equal

weights LR, and the Terwilliger's likelihood ratio with weights equal to p_j's TLR. For the score statistic we used the chi square distribution with one degree of freedom to approximate the distribution under the null hypothesis. For the LR and TLR statistics we used the 50:50 mixture of two chi squares with zero and one degree of freedom. We generated 10,000 samples of 200 case chromosomes and 200 control chromosomes from the multinomial distributions with probabilities $p_1 \cdots p_m$ for m equal to 4, 5, 8, 10, 15 and 20 haplotypes. Similar to the simulation described by Terwilliger, the frequency of the most common haplotype, p_1, was set to 0.5, whereas the remaining haplotypes were equally frequent ($0.5/(m-1)$). The results are shown in left columns of table 6.1.

For all m, the type I error rates of the score statistic $\hat{S}_{\frac{1}{m}}$ were maintained at the nominal error rate. For $m < 10$, the type I error rates of Pearson's chi square corresponded to the nominal level. However for larger m the type I error rates became conservative due to sparse data. For all m considered, the type I error rates for the TLR statistic were conservative (< 0.03). The type I error rates for the LR statistic were also somewhat small (≈ 0.04), but were better than the type I error rates for the TLR statistic.

To evaluate the power of the statistics, we generated 10,000 samples of n_1 case chromosomes and n_2 control chromosomes from the multinomial distributions with probabilities $p_1(1-\lambda)+\lambda$ and $p_j(1-\lambda)$ for $j = 2 \cdots m$ for cases and p_j $j = 1 \cdots m$ for controls, respectively. First, we considered the model used by Terwilliger (1995). The most common haplotype frequency p_1 in controls was again set to 0.5 and this haplotype was more frequent in cases. The parameter λ was fixed to 0.5 which corresponds to a haplotype frequency of 0.75 in the cases. The number of haplotypes m was again set to 4, 5, 8, 10, 15 and 20. The number of case and control chromosomes, n_1 and n_2 were now 100. The results are shown in the right columns of table 6.1.

For $m < 8$, the power of Pearson chi square was good. For $m \leq 8$ the power of the score statistic $\hat{S}_{\frac{1}{m}}$ was similar to the power of TLR statistic, while for $m > 8$, the TLR statistic had higher power than $\hat{S}_{\frac{1}{m}}$. For $m > 8$, the haplotype frequencies of the non associated haplotypes become too small yielding a large variance of the score statistic (see formula 6.3). The LR statistic appeared to perform worse than both $\hat{S}_{\frac{1}{m}}$ and TLR. Therefore we did not consider this statistic in the following simulations.

Second, we studied the power of the Pearson's chi square, $\hat{S}_{\frac{1}{m}}$, and TRL as function of the excess frequency λ for various values of the frequency of the associated haplotype $p_1 = 0.1, 0.2, 0.3, 0.4$ and 0.5. The remaining haplotypes

were again equally frequent. We restricted ourselves to a number of observed haplotypes m of 5 and 8, because most of the genes can be described by up to 8 common variants. The parameter λ was varied between 0 and 0.5. The number of chromosomes n_1 and n_2 were 200. We used a nominal significance level of 0.05. The results are depicted in figure 6.1.

For $p_1 = 0.5$, the score statistic $\hat{S}_{\frac{1}{m}}$ and likelihood ratio TLR performed similarly, and better than Pearson's chi square. For $m = 5$ and $p_1 = 0.4$ or 0.3 and for $m = 8$ and $p_1 = 0.4$, 0.3 or 0.2, the score statistic $\hat{S}_{\frac{1}{m}}$ performed better than TLR especially for small λ. For $p_1 = 0.2$ and $m = 5$ all three statistics had similar power. For $p_1 = 0.1$ and $m = 5$ or $m = 8$, Pearson's chi-square performs similar to TLR. Both statistics performed better than $\hat{S}_{\frac{1}{m}}$ except for$\lambda \leq 0.1$ and $p_1 = 0.1$ and $m = 5$. Note that for $p_1 = 0.1$ and $m = 5$, the power of the score statistic was small around $\lambda = 0.2$. This drop in power was due to the fact that the expectation of $\hat{S}_{\frac{1}{m}}$ becomes small (see formula 6.4).

Especially for common variants with frequency p_1 of 0.3 or 0.2 and a small impact on the disease ($\lambda \leq 0.1$), the score statistic performed well. For $m = 5$ and $m = 8$ and $p_1 = 0.3$, the gain in power of the score statistic compared to TLR statistic was about 4% and 8% for λ of 0.05 and 0.1 respectively. For $m = 5$ and $p_1 = 0.2$, both statistics performed similar. For $m = 8$ and $p_1 = 0.2$ the gain in power of the score statistic was large, namely 7% and 12% for λ of 0.05 and 0.1 respectively.

Data example

We applied the three statistics to a study on association between haplotypes of fibrinogen alpha (FGA), beta (FGB) and gamma (FGG) and the risk of deep venous thrombosis in LETS (Koster et al., 1993; van der Meer et al., 1997). Fifteen haplotype tagging SNPs were typed in 474 cases and 474 controls (Uitte de Willige et al., 2005). Within the three genes, the SNPs were in high linkage disequilibrium ($r_h^2 > 0.95$). The number of common haplotypes (frequency larger than 5%) describing FGG, FGA and FGB were three, five and five respectively. Since we focus on common haplotypes, we pooled the rare haplotypes with the less frequent haplotype category with frequency larger than 5%. In this analysis we considered p-values below 0.05 to be significant. In table 6.2 the data are described and the results are given.

For all genes, haplotype H2 appeared to be more frequent in the cases than in the controls. For FGG, FGA and FGB the allelic odds ratios of presence of H2 versus the rest was 1.34, 1.29 and 1.28 respectively. Note that these

odds ratios were rather similar while the p-values of the corresponding chi square statistics were different namely, 0.008, 0.051 and 0.059 respectively. The difference in p-values was caused by the difference in degrees of freedom of the chi square statistics and the frequencies of the other haplotypes. From the results of the standard chi-square statistics we concluded that only FGG was significantly associated to thrombophilia.

The p-values of the TLR were respectively 0.004, 0.021 and 0.078. These p-values were in line with the estimates of λ, namely 0.09, 0.07 and 0.05 respectively. Since FGA and FGB both had 5 variants, the frequency of the associated haplotype H2 was 0.3 and 0.2 respectively and the λ's were rather small, the score statistic should have more power than TLR for these genes.

The p-values of the score statistic $\hat{S}_{\frac{1}{m}}$ were 0.021, 0.007, 0.024 for FGG, FGA and FGB respectively. Indeed the p-values for FGA and FGB were smaller than the corresponding p-values of TLR statistic. The p-values for FGG and for FGB were larger than for FGA, because the frequencies of H2 in the combined case control sample were around $\frac{1}{m}$ (see formula 6.4). Based on the results of the score statistic, all genes were significantly associated to thrombophilia.

6.4 Discussion

In this report we have derived a new score statistic to test for association between a candidate gene and a binary trait. For candidate genes with a small impact on the disease and five to eight observed variants this new statistic appears to perform better than Terwilliger's LR statistic. Moreover the statistic is easy to compute and follows a chi square distribution under the null hypothesis. For more than eight variants, Terwilliger's LR statistic is more powerful. However by pooling less frequent haplotypes, the number of observed haplotypes is often smaller than eight.

Instead of using multi-locus haplotypes, some authors advocate to test each locus separately (Clayton et al., 2004). However, since mutations arise on haplotypes and because of the high degree of linkage disequilibrium, we prefer haplotype based tests for highly structured genes (Clark, 2004). For candidate genes that exists of several blocks we suggest to apply the test for each block separately. Alternatively the uncertainty has to be taken into account.

For multi allelic markers, Slager and Schaid (2001) and Czika and Weir (2004) proposed a multi allelic version of the trend test to test for association of genotypes. Also for genotypes at multi allelic markers, Houwing-Duistermaat

and Elston (2001) considered various ways to test for association using logistic regression models. If Hardy-Weinberg equilibrium does not hold, these methods should be preferred. Further, the parametrization used has a lower bound for the parameter λ which is often larger than -1. An alternative, more symmetrical parametrization might be the log relative risk corresponding to a logistic model. Moreover, in logistic models adjustments for other covariates are easily made. More research is needed to build these kind of models and derive corresponding tests for pairs of haplotypes.

We conclude that by choosing alternative weights, in particular constant weights, in the likelihood of Terwilliger, a set of new powerful and robust statistical tests was derived. For genetic association studies aiming to identify common associated variants, we recommend to first pool rare variants and then apply both the standard Pearson's chi square statistic as well as the new score statistic. By using both statistics more insight in the data can be obtained. A program is freely available which computes the statistics and corresponding p-values.

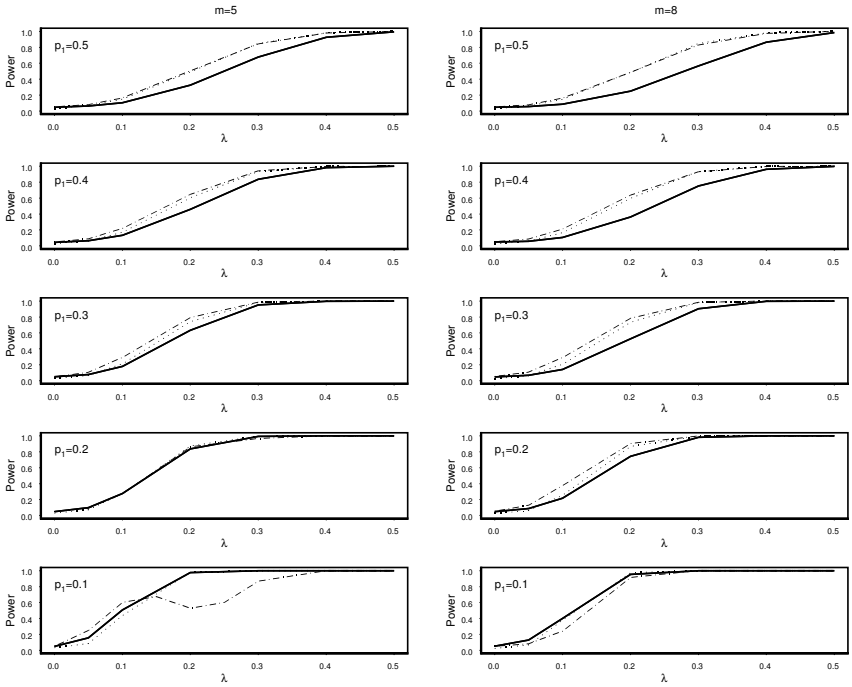

FIGURE 6.1: *Power curves of χ^2 (—), $\hat{S}_{\frac{1}{m}}$ (- - -) and TLR (·····) as function of λ for various values of the frequency of the associated variant $p_1 = 0.1, 0.2, 0.3, 0.4, 0.5$ and $m = 5$ (left column), and $m = 8$ (right column)*

TABLE 6.1: Type I error rate and power of the statistics χ^2 , $\hat{S}_{\frac{1}{m}}$, LR, TLR.

		type I error rate				power when $\lambda = 0.5$			
m	nominal	χ^2	$\hat{S}_{\frac{1}{m}}$	LR	TLR	χ^2	$\hat{S}_{\frac{1}{m}}$	LR	TLR
4	0.05	0.053	0.052	0.042	0.032	0.91	0.94	0.91	0.95
	0.01	0.009	0.010	0.010	0.007	0.78	0.84	0.80	0.86
	0.001	0.001	0.001	0.001	0.000	0.53	0.62	0.56	0.65
5	0.05	0.053	0.048	0.040	0.032	0.88	0.94	0.89	0.95
	0.01	0.010	0.010	0.010	0.007	0.71	0.83	0.77	0.87
	0.001	0.001	0.001	0.001	0.000	0.44	0.59	0.53	0.64
8	0.05	0.048	0.049	0.038	0.035	0.77	0.92	0.86	0.95
	0.01	0.008	0.010	0.008	0.004	0.53	0.80	0.74	0.86
	0.001	0.001	0.000	0.001	0.000	0.25	0.54	0.50	0.65
10	0.05	0.045	0.051	0.034	0.023	0.70	0.90	0.84	0.95
	0.01	0.006	0.008	0.008	0.003	0.43	0.76	0.70	0.85
	0.001	0.000	0.001	0.000	0.000	0.18	0.49	0.47	0.60
15	0.05	0.043	0.048	0.041	0.020	0.58	0.86	0.80	0.95
	0.01	0.007	0.010	0.010	0.003	0.31	0.69	0.65	0.85
	0.001	0.000	0.001	0.002	0.000	0.09	0.42	0.42	0.63
20	0.05	0.043	0.052	0.045	0.021	0.48	0.84	0.77	0.95
	0.01	0.005	0.011	0.010	0.004	0.20	0.64	0.62	0.84
	0.001	0.000	0.001	0.002	0.000	0.04	0.35	0.39	0.60

TABLE 6.2: Descriptives and results of genetic association on LETS.

haplotype	case chromosomes	control chromosomes	χ^2	$\hat{S}_{\frac{1}{m}}$	TLR	$\hat{\lambda}$
FGG ($n_1 = 938$, $n_2 = 942$)			0.008	0.021	0.004	0.09
H1	334 (36.3)	366 (38.9)				
H2	315 (33.2)	254 (27.0)				
H3+H4	289 (30.5)	321 (34.1)				
FGA ($n_1 = 936$, $n_2 = 942$)			0.051	0.007	0.021	0.07
H1	270 (28.8)	266 (28.2)				
H2	320 (34.2)	270 (28.7)				
H3	95 (10.2)	117 (12.4)				
H4	100 (10.7)	121 (12.9)				
H5	151 (16.1)	168 (17.8)				
FGB ($n_1 = 936$, $n_2 = 932$)			0.059	0.024	0.078	0.05
H1	328 (35.0)	310 (33.3)				
H2	231 (24.7)	189 (20.3)				
H4	135 (14.4)	149 (16.0)				
H6	128 (13.7)	143 (15.3)				
H3+H5+H7	114 (13.2)	141 (15.1)				